DOCUMENT RESUME

ED 462 422 TM 033 686

AUTHOR Childs, Ruth A.; Oppler, Scott H.

TITLE Practical Implications of Test Dimensionality for Item Response Theory Calibration of the Medical College Admission Test. MCAT Monograph.

INSTITUTION American Institutes for Research, Washington, DC.
SPONS AGENCY Association of American Medical Colleges, Washington, DC.
REPORT NO MCAT-1
PUB DATE 1999-07-30
NOTE 43p.
AVAILABLE FROM Association of American Medical Colleges, Section for the Medical College Admission Test, 2450 N Street, NW, Washington, DC 20037. Tel: 202-828-0400; Fax: 202-828-1125; Web site: http://www.aamc.org/mcat.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Entrance Examinations; Higher Education; *Item Response Theory; Medical Education; *Scoring
IDENTIFIERS Calibration; *Dimensionality (Tests); *Medical College Admission Test

ABSTRACT

The use of item response theory (IRT) in the Medical College Admission Test (MCAT) testing program has been limited. This study provides a basis for future IRT analyses of the MCAT by exploring the dimensionality of each of the MCAT's three multiple-choice test sections (Verbal Reasoning, Physical Sciences, and Biological Sciences) and the implications of that dimensionality for IRT calibration and scoring. Three sets of data were used: an MCAT form administered on 2 occasions (16,520 and 3,638 observations) and a second, content-equivalent form administered on the second occasion (12,625 observations). Dimensionality analyses suggested that while the items within each of the science test sections are not completely homogeneous, nor are they measuring distinct constructs corresponding to the disciplines. As is consistent with this finding, when confirmatory factor analyses were performed, the Biological Sciences disciplines were found to be correlated about 0.85, and the Physical Sciences about 0.95. Comparison of the results of the IRT calibration and scoring of the data by test and by discipline within the test suggests that the tests do not deviate sufficiently from unidimensionality to require that the IRT item calibrations be performed separately by discipline within the test. The use of these results to inform future IRT analyses of the MCAT depends in large part on policy decisions about future directions of the MCAT program. (Contains 17 tables, 3 figures, and 20 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

ASSOCIATION OF AMERICAN MEDICAL COLLEGES

MCAT
MEDICAL COLLEGE ADMISSION TEST

ED 462 422

MCAT Monograph 1

Practical Implications of Test Dimensionality for IRT Calibration of the MCAT

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
P. Etienne
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033686

BEST COPY AVAILABLE

1

July 1999

PRACTICAL IMPLICATIONS OF TEST DIMENSIONALITY FOR ITEM RESPONSE THEORY CALIBRATION OF THE MEDICAL COLLEGE ADMISSION TEST

Ruth A. Childs & Scott H. Oppler
American Institutes for Research
Washington, DC

July 30, 1999

For additional printed copies of this Monograph contact the MCAT Section at AAMC, (202) 828-0690

©1999, Association of American Medical Colleges. All materials subject to this copyright may be photocopied for the non-commercial purposes of scientific or educational advancement.

Printed in the USA.

EXECUTIVE SUMMARY

The previous use of item response theory (IRT) in the Medical College Admission Test (MCAT) testing program has been limited. This study provides a basis for future IRT analyses of the MCAT by exploring the dimensionality of each of the MCAT's three multiple-choice test sections (Verbal Reasoning, Physical Sciences, and Biological Sciences) and the implications of that dimensionality for IRT calibration and scoring.

Three sets of data were used in this study: an MCAT form administered on two occasions (16,520 and 3,638 observations) and a second, content-equivalent form administered on the second occasion (12,625 observations).

DIMENSIONALITY ANALYSES

Each of the three MCAT multiple-choice test sections is scored separately, so that candidates receive one score for Verbal Reasoning, one score for Physical Sciences, and one score for Biological Sciences. However, each of the science sections measures knowledge of two distinct disciplines. The Physical Sciences test section includes Physics and General Chemistry items and the Biological Sciences section includes Biology and Organic Chemistry items. This study focuses on the discipline-related dimensionality within each of the MCAT test sections.

Three methods were used to explore the dimensionality of the test sections. Stout's Test of Essential Unidimensionality (implemented in the computer program DIMTEST) suggested that the test sections -- and the disciplines within the science sections – were not strictly unidimensional. However, DIMTEST checks for departure from unidimensionality and does not indicate the degree to which the tests are nonunidimensional and the practical implications of the nonunidimensionality.

Nonlinear factor analyses (implemented in NOHARM) of the data were performed in both exploratory and confirmatory modes. The exploratory analyses for the Biological Sciences test section found two factors corresponding roughly to the discipline categories. For the Physical Sciences section, however, the exploratory analyses seeking a two-factor solution yielded factors unrelated to the discipline categories. When confirmatory NOHARM analyses were performed, with the discipline categories defining the factors, the Biological Sciences discipline factors were found to be correlated .77 to .81 and the Physical Sciences discipline factors were correlated .90 to .97.

Confirmatory factor analyses, based on tetrachoric correlations among the items within each test section, were also performed using LISREL. The LISREL analyses compared the fits of the discipline-defined factor solutions and a single-factor solution within each test. Although the χ^2 difference tests suggested a better fit for the two-factor model, for each combination of test section, form, and administration, the Root Mean Square Error of Approximation (RMSEA) indices were very similar across the models, suggesting that the one- and two-factor models fit the data about equally well. The correlations between the factors defined by the Biological Sciences disciplines were .81 to .91; between the factors defined by the Physical Sciences disciplines, about .94 to .97.

IRT ANALYSES

Following the dimensionality analyses, a series of IRT calibrations of the items in the data sets were also performed. The one-, two-, and three-parameter logistic (1PL, 2PL, and 3PL) models were applied to each test and to each discipline within test. Comparisons were made between score estimates based on the 1PL, 2PL, and 3PL models and the original raw and scale scores. In addition, the stability of the item parameter estimates across administrations was examined for the form that had been administered twice. Parallel sets of analyses were performed with unanswered items counted as incorrect or omitted from the analyses.

The results of the IRT analyses were similar for all data sets, regardless of the treatment of answered items. The 3PL fit the data slightly better than the 1PL and 2PL models and provided greater score precision at the higher scores, while the 1PL model provided greater precision at the lower scores. The difficulty (*b*) parameters were very stable across the two repeated administrations for all three models, the slope (*a*) parameters were slightly less so, and the asymptote (*c*) parameters were the least stable. For the science sections, the item parameter estimates were relatively unaffected by whether they were estimated within or across disciplines, suggesting that the discipline-based multidimensionality of the science sections may be immaterial in the calibration of the MCAT item bank.

The correlations among the various types of scores (raw, scale, and IRT ability estimates) were very high (above .97) and the correlations across the tests and between the disciplines within the science test sections were very similar for all of the score types. In addition, for both the science sections, the relative ordering of examinees produced by combining the separate scores based on the separate discipline calibrations was almost identical to the ordering based on a single calibration.

SUMMARY

The dimensionality analyses suggested that, while the items within each of the science test sections are not completely homogeneous, neither are they clearly measuring distinct constructs corresponding to the disciplines. While the Biological Sciences test section showed some evidence of two factors corresponding to the disciplines, the Physical Sciences section did not appear to support two distinguishable factors. Consistent with this finding, when confirmatory factor analyses were performed, the Biological Sciences disciplines were found to be correlated about .85; the Physical Sciences disciplines, about .95.

The motivation for performing these analyses was to determine whether the tests deviated sufficiently from unidimensionality to require that IRT item calibrations be performed separately by discipline within test. Comparisons of the results of the IRT calibration and scoring of the data by test and by discipline within test suggested that they do not. For example, the correlations between IRT scores based on item parameters estimated separately by discipline and formed into weighted composites and scores based on item parameters that were estimated across disciple within test exceeded .99.

The use of these results to inform future IRT analyses of the MCAT depends in large part on policy decisions about future directions of the MCAT Program. However, these results may also have implications for other testing programs as those programs seek to determine the practical implications of dimensionality for the IRT calibration of their tests.

CONTENTS

1. Introduction 1
   About the MCAT 1
   MCAT Data Used in This Study 2
2. Dimensionality Analyses 3
   Approaches to Investigating Dimensionality 4
   Stout's Test of Essential Unidimensionality Analyses (DIMTEST) 5
   Non-Linear Factor Analyses (NOHARM) 7
   Linear Factor Analyses (LISREL) 8
   Summary 9
3. IRT Calibration 10
   Item Response Theory Models 10
   Calibration and Scoring Procedures 11
   Analysis Results 12
      Model Fit 12
      Item Parameter Estimate Stability Across Administrations 14
      Influence of Discipline-Based Multidimensionality on Item Parameter Estimates for Science Test Sections 15
      Score Comparability Within Test Section 16
      Influence of Discipline-Based Multidimensionality on Score Estimates for Science Test Sections 20
      Score Comparability Across Test Sections 20
      Score Comparability Across Disciplines 21
   Summary 22
4. Treatment of Missing Data 24
5. Conclusions 30

References 31

TABLES

Table 1. MCAT Forms Selected for This Study 2
Table 2. Percents of Items of Each Discipline Within Test Section in Selected Forms 3
Table 3. Some Approaches to Investigating Dimensionality 4
Table 4a. Results of Principal DIMTEST Analyses 6
Table 4b. Results of Supplementary DIMTEST Analyses Using Only 2,000 Observations in Computation of *T* (Form X in 1994) 6
Table 4c. Results of Supplementary DIMTEST Analyses By Discipline Within Test Section (Form X in 1994) 7
Table 5a. Results of Exploratory NOHARM Analyses 8
Table 5b. Results of Confirmatory NOHARM Analyses 8
Table 6. Results of LISREL Analyses 9
Table 7. Model Comparisons for MULTILOG Item Parameter Estimation Assuming Each Test Section to be Unidimensional 12
Table 8. Correlations Among Item Parameter Estimates: Form X in 1994 and Form X in 1996 14
Table 9. Correlations Among Item Parameter Estimates Calibrated by Test Section Across Discipline and Within Discipline 16
Table 10a. Correlations Among Raw, Scale, and IRT-Based Scores: Form X in 1994 and 1996, Verbal Reasoning Test Section 17
Table 10b. Correlations Among Raw, Scale, and IRT-Based Scores: Form X in 1994 and 1996, Physical Sciences Test Section 17
Table 10c. Correlations Among Raw, Scale, and IRT-Based Scores: Form X in 1994 and 1996, Biological Sciences Test Section 17
Table 11a. Correlations Among Raw, Scale, and IRT-Based Scores: Form Y in 1996, Verbal Reasoning Test Section 17
Table 11b. Correlations Among Raw, Scale, and IRT-Based Scores: Form Y in 1996, Physical Sciences Test Section 18
Table 11c. Correlations Among Raw, Scale, and IRT-Based Scores: Form Y in 1996, Biological Sciences Test Section 18
Table 12. Correlations Among IRT-Based Score Estimates for Science Test Section Scores Based on Within- and Across-Discipline Item Calibrations 20
Table 13. Correlations Among Test Sections for Scale and IRT-Based Score Estimates 21
Table 14. Correlations Among Disciplines for IRT-Based Score Estimates 21
Table 15. Numbers of Items Omitted or Not Reached 25

Table 16. Correlations Among Item Parameter Estimates Across Method of Scoring Missing Data: Missing Items Ignored and Missing Items Scored as Wrong 26
Table 17. Correlations Among IRT-Based Score Estimates Across Method of Scoring Missing Data: Missing Items Ignored and Missing Items Scored as Wrong 27

FIGURES

Figure 1a. Test Information Curves: Form X in 1994, Verbal Reasoning Test Section 13

Figure 1b. Test Information Curves: Form X in 1994, Physical Sciences Test Section 13

Figure 1c. Test Information Curves: Form X in 1994, Biological Sciences Test Section 14

Figure 2a. Examinee Score Estimate Scatter Plots: Form X, 1994, Biological Sciences Test section 1-Parameter Logistic Model and Raw Scores 18

Figure 2b. Examinee Score Estimate Scatter Plots: Form X, 1994, Biological Sciences Test Section, 2-Parameter Logistic Model and Raw Scores 19

Figure 2c. Examinee Score Estimate Scatter Plots: Form X, 1994, Biological Sciences Test Section, 3-Parameter Logistic Model and Raw Scores 19

Figure 3a. Examinee Score Estimate Scatter Plots: Form X, 1994, Verbal Reasoning Test Section, 1-Parameter Logistic Model, Missing Items Ignored and Missing Items Scored as Wrong 27

Figure 3b. Examinee Score Estimate Scatter Plots: Form X, 1994, Verbal Reasoning Test Section, 2-Parameter Logistic Model, Missing Items Ignored and Missing Items Scored as Wrong 28

Figure 3c. Examinee Score Estimate Scatter Plots: Form X, 1994, Verbal Reasoning Test Section, 3-Parameter Logistic Model, Missing Items Ignored and Missing Items Scored as Wrong 28

ACKNOWLEDGEMENTS

In 1997 and early 1998, the American Institutes for Research (AIR), at the request of the Association of American Medical Colleges (AAMC), performed a series of analyses to establish a basis for the possible future item response theory (IRT) calibration of the Medical College Admission Test (MCAT) item bank. This report summarizes the results of those analyses.

The authors are grateful to the director and staff of the AAMC's MCAT Program and to the Program's advisors and consultants for their helpful suggestions throughout this study. In particular, MCAT program director Ellen Julian, program staff members Judith Koenig and Kristen Huff, advisors Linda Crocker (University of Florida at Gainesville), Richard Jaeger (University of North Carolina at Greensboro), and Robert Linn (University of Colorado at Boulder), and consultant Steve Sireci (University of Massachusetts at Amherst) provided many valuable comments. Also, without the tactful suggestions and expert assistance of the AAMC's MCAT Program staff and Publication Department staff, the transformation of the original voluminous report into this monograph would not have been accomplished.

1. INTRODUCTION

Item response theory (IRT) analyses have added much to our understanding of the relationships among and characteristics of test items, as revealed in examinees' response patterns. Use of IRT in large-scale assessment programs has increased dramatically over the past two decades.

Because the application of IRT in the Medical College Admission Test (MCAT) testing program has so far been limited, this study is designed to provide a basis for possible future IRT analyses of the MCAT. This report describes a detailed investigation of the dimensionality of the MCAT's three multiple-choice test sections and of the practical implications of the dimensionality analysis findings with regard to IRT calibration and scoring.

This chapter describes the MCAT and the MCAT data used in this study. Chapter 2 details the dimensionality analyses, and Chapter 3 describes the IRT calibrations and score estimations. Chapter 4 compares the results of two approaches to treating missing data. The final chapter, Chapter 5, summarizes the study's findings.

ABOUT THE MCAT

The Association of American Medical Colleges (AAMC) provides the MCAT as an admissions tool for its member medical schools. The MCAT is administered twice annually. At each administration, approximately 30,000 candidates for admission to medical school take the exam.

The MCAT consists of four paper-and-pencil test sections. Three test sections (Verbal Reasoning, Physical Sciences, and Biological Sciences) consist of multiple-choice items. A fourth test section requires candidates to write two brief essays on assigned topics. As noted above, this study focuses on the three multiple-choice test sections.

The Verbal Reasoning test section contains nine reading passages (500 to 600 words in length), each with six to ten associated items. The science test sections (Physical Sciences and Biological Sciences) each have eleven problem sets of four to seven items, plus fifteen items that are not linked to problem sets.

MCAT DATA USED IN THIS STUDY

Two MCAT test forms were selected for use in this study. One of the test forms was administered twice, in administrations two years apart. The selected test forms -- referred to throughout this study as X and Y, the administration dates, and the number of usable observations for each are shown in Table 1.

TABLE 1. MCAT FORMS SELECTED FOR THIS STUDY

Form	Administration	Number of Observations
X	1994	16,520
X	1996	3,638
Y	1996	12,625

For some of the computations (e.g., the computation of tetrachoric correlations for the dimensionality analyses), examinees who omitted items were excluded. Because the number of excluded observations varied by analysis, all observations are included in Table 1.

2. DIMENSIONALITY ANALYSES

The MCAT is developed with the assumption that each test section -- Verbal Reasoning, Biological Sciences, and Physical Sciences -- represents a different ability dimension, so that the multidimensionality of concern in this study is that within test sections, not across them. Two of the three test sections -- Biological Sciences and Physical Sciences – each contain items from two distinct disciplines, intended to measure knowledge in different parts of the domain defined by the test specifications. Specifically, the Biological Sciences test section contains items that are categorized as either Biology or Organic Chemistry items, and the Physical Sciences test section contains both Physics and General Chemistry items. The percent of items categorized in each discipline for each form analyzed in this study are reported in Table 2.

TABLE 2. PERCENTS OF ITEMS OF EACH DISCIPLINE WITHIN TEST SECTION IN SELECTED FORMS

Form	Test Section	Discipline	Percent of Test Section Items
X	Verbal Reasoning	n/a	100
	Physical Sciences	Physics	48
		General Chemistry	52
	Biological Sciences	Biology	68
		Organic Chemistry	32
Y	Verbal Reasoning	n/a	100
	Physical Sciences	Physics	51
		General Chemistry	49
	Biological Sciences	Biology	70
		Organic Chemistry	30

The dimensionality analyses in this study focus on whether the two disciplines within each of the science test sections represent distinct dimensions. It is possible that other subsets of items may also form distinct dimensions for the purposes of IRT item calibration; however, multidimensionality related to the disciplines was judged most likely to have practical implications for future IRT calibrations. Of course, in addition to determining whether a test section may have more than one dimension, it is important to investigate to what degree the dimensions are distinct from one another and to consider at what degree of distinctness it becomes essential to incorporate this information into the IRT item calibration design.

APPROACHES TO INVESTIGATING DIMENSIONALITY

A number of methods are available for investigating dimensionality in a set of items. Table 3 summarizes the major advantages and disadvantages of the methods that were considered for this study.

TABLE 3. SOME APPROACHES TO INVESTIGATING DIMENSIONALITY

Approach	Advantages	Disadvantages
Item-Level Linear Factor Analysis of Tetrachoric Correlations (e.g., TESTFACT)	Traditional	Fairly high type I error rates (rejects unidimensionality when the data are in fact unidimensional; De Champlain & Gessaroli, in press)
Item-Level Nonlinear Factor Analysis (e.g., NOHARM)	Gessaroli and De Champlain's (1996) approximate chi-square test for use with NOHARM shows low type I errors and good statistical power (Gessaroli & De Champlain, 1996); can be used to determine the actual dimensionality of multidimensional data	May not identify highly correlated dimensions in multidimensional data (Nandakumar, 1994)
Multidimensional Scaling (e.g., ALSCAL)	Easy to use; has conceptual similarity to IRT in relating items to an ability scale (DeAyala & Hertzog, 1991)	Generally only useful with ordinal polytomous data, although new measures of item proximity may help overcome this limitation (DeAyala & Hertzog, 1991)
Confirmatory Factor Analysis (e.g., LISREL, LISCOMP)	Tests a priori dimensions	Requires large data sets (e.g., 1,000 examinees per item; De Champlain & Gessaroli, in press)
Stout's Test of Essential Unidimensionality (i.e., DIMTEST)	Low type I errors and good statistical power, except for small sample sizes or small item sets (Gessaroli & De Champlain, 1996)	Only tests for unidimensionality, doesn't determine actual number of dimensions; may underestimate dimensionality if guessing is prevalent (Hattie, Krakowski, Rogers, & Swaminathan, 1996)
IRT Parameter Comparison (Bejar, 1980)	Conceptually consistent with subsequent IRT analyses (Bejar, 1980)	May not detect multidimensionality (Hambleton & Rovenelli, 1986)
IRT Residual Analysis	Conceptually consistent with subsequent IRT analyses	May not detect multidimensionality (Hambleton & Rovenelli, 1986)
Holland-Rosenbaum Procedure	Conceptually consistent with subsequent IRT analyses	Very conservative (accepts unidimensionality when the data are in fact multidimensional; Nandakumar, 1994)

Note: This table is based, in part, on descriptions of methods for investigating dimensionality in Sireci (1997).

Following a review of these methods for investigating dimensionality, three methods were selected for the dimensionality analyses in this study:

1. Stout's Test of Essential Unidimensionality (as implemented in DIMTEST);
2. Item-level nonlinear factor analysis (as implemented in NOHARM); and
3. Confirmatory factor analysis (as implemented in LISREL).

The results of these analyses are described in the following sections.

STOUT'S TEST OF ESSENTIAL UNIDIMENSIONALITY ANALYSES (DIMTEST)

Stout's Test of Essential Unidimensionality was applied using the program DIMTEST (Stout, 1987; Stout, Douglas, Junker, & Roussos, 1993). The program involves three steps:

1. Selecting, based either on *a priori* hypotheses or factor analysis results, a set of items (fewer than 1/3 of the total items) that is expected to be unidimensional -- this is referred to as Assessment Test 1 or AT1;

2. Running a program component that selects a second set of items that are matched in difficulty with the first set -- this is called Assessment Test 2 or AT2; and

3. Running a second program component that scores the examinees' responses to the remaining items -- the Partitioning Test or PT -- and divides them into groups on the basis of their scores. It then compares statistics for each group of examinees based on their variances on AT1 and AT2.

According to Stout et al. (1993), "Functionally, each run of the procedure assesses whether the chosen AT1 is dimensionally distinct from PT with the chosen AT2 to be used to eliminate the statistical biasing effect on the procedure that would otherwise result from the fact that examinees are matched in subgroups using an unreliable observed score PT" (p. 2). The statistic resulting from these computations is *T*.

The DIMTEST analyses for the test sections were conducted in two ways:

1. The internal factor analysis option in DIMTEST was allowed to select AT1 and the rest of the analysis was based on this empirical selection, or

2. AT1 was specified *a priori* as the smaller of the two sets of items within each test section.

Because DIMTEST specifies that AT1 must contain fewer than 1/3 of the items in the test section, option 2 was applied only for the Biological Sciences test section, which contains slightly fewer than 1/3 Organic Chemistry items (the rest being Biology items); for the Physical Sciences test section, the percentages of Physics and General Chemistry items are almost equal, so only option 1 was applied. When the factor analysis procedure was applied, it was performed on a randomly selected half of the examinees and the DIMTEST statistic was computed using the remaining examinees.

For the analyses described in this report, the version of DIMTEST updated in 1993 was used. It should be noted, however, that this version of DIMTEST seems to be overly sensitive to sample size (Stout, personal communication, June 18, 1997). A version of DIMTEST released after these analyses were completed may improve the power of the analysis.

To investigate the hypothesis that the large numbers of examinees being used in these computations were responsible for the levels of significance found, a set of analyses for the Form X 1994 data was also run, this time using a random sample of only 2,000 examinees in the computation of *T*, the DIMTEST statistic. In addition, analyses using all of the Form X 1994 data were run within discipline for the science test sections to investigate how the DIMTEST statistic would behave for smaller, possibly more homogeneous sets of items.

Tables 4a, 4b, and 4c summarize the results of the DIMTEST analyses. The significance level suggests whether or not to reject the null hypothesis that the test sections are each unidimensional. With only one exception, the results of the principal DIMTEST analyses reported in Table 4a indicate significant departure of the data from unidimensionality, regardless of form, test section, administration, or AT1 selection method. If one takes into account the large number of significance tests being performed, only one additional combination fails to reach significance. Tables 4b and 4c show similar results, even though the numbers of examinees have been reduced in Table 4b, and the homogeneity of the items has been increased in Table 4c. Taken at face value, these results would suggest that each test section -- and, indeed, each discipline within each test section -- is not unidimensional. However, as noted at the beginning of this chapter, answering a simple "yes" or "no" to the question, "are the test sections unidimensional?" is not sufficient. The following questions must also be answered: "to what degree are they nonunidimensional?" and "what are the practical implications of that degree of nonunidimensionality?" The non-linear and linear factor analyses reported later in this chapter are intended to help answer these questions.

TABLE 4A. RESULTS OF PRINCIPAL DIMTEST ANALYSES

Form	Test Section	AT1 Selection Method	Administration	*n*	*T*	*p*
X	Verbal Reasoning	Factor Analysis	1994	7,814	13.47	<.0001
			1996	1,732	9.53	<.0001
	Physical Sciences	Factor Analysis	1994	7,004	6.13	<.0001
			1996	1,739	1.16	.12
	Biological Sciences	Factor Analysis	1994	7,740	10.72	<.0001
			1996	1,673	2.66	<.01
		Organic Chemistry Items	1994	16,157	11.54	<.0001
			1996	3,587	5.23	<.0001
Y	Verbal Reasoning	Factor Analysis	1996	5,902	9.96	<.0001
	Physical Sciences	Factor Analysis	1996	5,940	8.33	<.0001
	Biological Sciences	Factor Analysis	1996	5,965	9.27	<.0001
		Organic Chemistry Items	1996	11,880	10.96	<.0001

Note: n = number of observations used in the computation of the T statistic after deletion by the DIMTEST program of some cases with missing data (when factor analysis was used to select AT1, the factor analysis was performed on a randomly selected half of the observations and the computation of the T statistic was performed on the remaining observations); T = DIMTEST statistic (conservative); p = observed significance.

TABLE 4B. RESULTS OF SUPPLEMENTARY DIMTEST ANALYSES USING ONLY 2,000 OBSERVATIONS IN COMPUTATION OF *T* (FORM X IN 1994)

Form	Test Section	AT1 Selection Method	Administration	*n*	*T*	*p*
X	Verbal Reasoning	Factor Analysis	1994	1,924	8.25	<.0001
	Physical Sciences	Factor Analysis	1994	1,922	1.79	<.05
	Biological Sciences	Factor Analysis	1994	1,908	5.34	<.0001

Note: n = number of observations used in the computation of the T statistic after deletion by the DIMTEST program of some cases with missing data (the factor analysis was performed on the 14,520 observations remaining after the 2,000 examinees for the computation of the T statistic were selected); T = DIMTEST statistic (conservative); p = observed significance.

TABLE 4C. RESULTS OF SUPPLEMENTARY DIMTEST ANALYSES BY DISCIPLINE WITHIN TEST SECTION (FORM X IN 1994)

Form	Test Section	Discipline	AT1 Selection Method	Administration	*n*	*T*	*p*
X	Physical Sciences	General Chemistry	Factor Analysis	1994	7,963	1.70	<.05
		Physics	Factor Analysis	1994	8,045	7.07	<.0001
	Biological Sciences	Biology	Factor Analysis	1994	7,864	8.22	<.0001
		Organic Chemistry	Factor Analysis	1994	6,831	3.57	<.01

Note: n = number of observations used in the computation of the T statistic after deletion by the DIMTEST program of some cases with missing data (when factor analysis was used to select AT1, the factor analysis was performed on a randomly selected half of the observations and the computation of the T statistic was performed on the remaining observations); T = DIMTEST statistic (conservative); p = observed significance.

NON-LINEAR FACTOR ANALYSES (NOHARM)

A second investigation of the dimensionality of the test sections was performed using the NOHARM program (Fraser, 1988), which performs non-linear factor analysis based on McDonald's approach (see, for example, McDonald, 1994). The NOHARM program was used in both exploratory and confirmatory modes, with the exploratory analyses seeking a two-factor solution for each test section and the confirmatory analyses performed for both one- and two-factor solutions.

Exploratory non-linear factor analyses seeking two factors were performed for each test section for each form and administration. The percentages of items with loadings greater than .3 or less than -.3 on one and only one factor in the factor loading matrix (after promax rotation) were calculated. The percentages of items for which these loadings were on the factor corresponding to their intended discipline (or in the case of Verbal Reasoning, the percentage of items loading on the first factor) were also calculated.

Confirmatory non-linear factor analyses for one- and two-factor solutions were performed for the Physical Sciences and Biological Sciences test sections. For the two-factor solutions, items were forced to load on the factor corresponding to the discipline in which they have been categorized (e.g., in the case of Physical Sciences, items were specified as loading on a factor corresponding to either Physics or General Chemistry). Gessaroli and De Champlain's (1996) approximate χ^2 statistic, implemented in the CHIDIM program (De Champlain & Tang, 1997), was used to assess the fit of these models. Following the procedure used by De Champlain and Gessaroli (in press), the dimensionality of the data should be assessed by examining the size of the approximate χ^2 statistic for a one-factor model fit to the data. However, as Gessaroli and De Champlain (1996) note, "this χ^2 statistic has the same limitations as other χ^2 indexes...one might expect to often falsely reject the correct m-factor model with large sample sizes..." (p. 160).

The results of these analyses are presented in Tables 5a and 5b. The results of the exploratory analysis, shown in Table 5a, suggest that the Verbal Reasoning and Biology test sections have factor structures that correspond roughly to the discipline categories. In other words, an examinee's success in answering an item is more highly related to his or her success in answering other items within that discipline than items in the other discipline within the test section. For the Verbal Reasoning test section, there is, of course, only one discipline, and most of the items loaded on the first factor. For the Physical Sciences test section, in contrast, the items were split across the two factors, but not in any way that related to the discipline categories.

The results of the confirmatory analyses, shown in Table 5b, indicate that factors defined by the discipline categories are quite highly correlated for the Physical Sciences test section (r = .90-.97); but less so for the Biological Sciences test section (r = .77-.81). The approximate χ^2 statistics suggest that the models do not fit well. However, as noted above, the large sample size makes interpretation of these statistics problematic.

Taken together, the results of the exploratory and confirmatory analyses suggest that the Biological Sciences test section does support moderately correlated factors that correspond roughly to the discipline categories, but that the Physical Sciences test section does not.

TABLE 5A. RESULTS OF EXPLORATORY NOHARM ANALYSES

Form	Test Section	Administration	n	Items in Test Section	r	One Loading < -.3 or > .3	Loading on Discipline
X	Verbal Reasoning	1994	14,138	55	.69	82%	71%
		1996	3,184	55	.71	85%	73%
	Physical Sciences	1994	14,035	63	.69	84%	43%
		1996	3,124	63	.66	78%	43%
	Biological Sciences	1994	15,694	63	.71	86%	81%
		1996	3,487	63	.70	76%	68%
Y	Verbal Reasoning	1996	11,171	55	.77	78%	51%
	Physical Sciences	1996	11,378	63	.74	71%	48%
	Biological Sciences	1996	11,663	63	.64	87%	87%

Note: n = number of observations used in NOHARM analysis after listwise deletion of any observations with missing data;
r = correlation between factors.

TABLE 5B. RESULTS OF CONFIRMATORY NOHARM ANALYSES

Form	Test Section	Administration	n	Items in Test Section	1 Dimension	2 Dimensions	
					χ^2	χ^2	r
X	Physical Sciences	1994	15,694	63	23,161	22,869	.96
		1996	3,124	63	6,307	6,312	.97
	Biological Sciences	1994	15,694	63	14,705	11,635	.78
		1996	3,487	63	5,761	5,274	.81
Y	Physical Sciences	1996	11,378	63	8,890	8,289	.90
	Biological Sciences	1996	11,663	63	16,164	13,384	.77

Note: n = number of observations used in NOHARM analysis after listwise deletion of any observations with missing data,
χ^2 for one dimension has 1890 degrees of freedom; for two dimensions, χ^2 has 1889 degrees of freedom;
r = correlation between factors.

LINEAR FACTOR ANALYSES (LISREL)

Confirmatory factor analyses, based on tetrachoric correlations among the items within each test section, were performed using LISREL to determine whether one or two factors best fit the items within each of the two science test sections. Specifically, for the Physical Sciences test section, the one-factor model was compared to a two-factor model in which Physics items were assigned to one factor, and General Chemistry items were assigned to the other. Likewise, the one-factor model for the Biological Sciences test was compared to a two-factor model splitting Biology and Organic Chemistry items.

In the first step, LISREL's data preparation package, PRELIS (version 2; Jöreskog & Sörbom, 1988, 1993b), was used to compute the tetrachoric correlations among the items within each of the test sections. (Because the items are scored dichotomously, the Pearson product-moment correlations tend to be underestimates of the correlations; tetrachoric correlations are estimates of correlations among hypothetical multivariate normal variables underlying the observed discrete responses [Mislevy, 1986].) Observations with any missing data were not included in this analysis. (This is necessary for the calculation of the tetrachoric correlations in PRELIS.)

In the second step, LISREL (version 8; Jöreskog & Sörbom, 1989, 1993a, 1993c) was used to test the fit of the models defined by the *a priori* sets of items within the test sections. As described above, confirmatory factor analyses specifying one factor and two factors were performed for each test section, using weighted least squares estimation and the tetrachoric correlations output by PRELIS.

The differences between the χ^2 statistics for the two models were computed and are reported in Table 6. In addition, the Root Mean Square Error of Approximation (RMSEA), a measure of misfit, is reported for each comparison. The RMSEA seems to be more sensitive to real differences in fit between models than the χ^2 statistic; a smaller RMSEA indicates better fit. Although the χ^2 difference tests suggest that the 2-factor model fits the data significantly better than the one-factor model for each combination of form, test section, and administration, the RMSEAs are virtually identical across models, suggesting that the one- and two-factor models fit about equally well. As noted above, the large sample sizes make interpretation of the χ^2 tests problematic, so the RMSEAs may be particularly worth considering for these comparisons. The other results summarized in Table 6 are also in line with the results of the NOHARM analyses -- the factors corresponding to the disciplines are highly correlated for both test sections, but the correlations are higher for the two Physical Sciences disciplines than for the two disciplines comprising the Biological Sciences test section.

TABLE 6. RESULTS OF LISREL ANALYSES

Form	Test Section	Administration	*n*	Items in Test Section	1 Dimension		2 Dimensions			Difference	
					χ^2	RMSEA	χ^2	RMSEA	*r*	χ^2	*p*
X	Physical Sciences	1994	14,035	63	10,803	.018	10,426	.018	.95	377	<.001
		1996	3,124	63	10,367	.038	10,252	.038	.97	115	<.001
	Biological Sciences	1994	15,694	63	10,633	.017	9,651	.016	.85	982	<.001
		1996	3,487	63	8,745	.032	8,483	.032	.91	262	<.001
Y	Physical Sciences	1996	11,378	63	7,398	.016	7,063	.016	.94	335	<.001
	Biological Sciences	1996	11,663	63	8,896	.018	7,640	.016	.81	1,256	<.001

Note: χ^2 for one dimension has 1890 degrees of freedom; for two dimensions, χ^2 has 1889 degrees of freedom; the test of the difference has one degree of freedom; RMSEA = Root Mean Square Error of Approximation; *r* = correlation between factors; *p* = significance of difference between fits of models (for χ^2 statistics).

SUMMARY

The results of these dimensionality analyses suggest some degree of multidimensionality in the test sections. However, the dimensions corresponding to the disciplines within test section are quite highly correlated. The IRT analysis results described in the following chapter provide additional information about the practical dimensionality of the test sections. The final chapter discusses the implications of these findings, taken together.

3. IRT CALIBRATION

Following the dimensionality analyses, a series of IRT calibrations of the items in the data sets was performed. As described below, the 1-, 2-, and 3-parameter logistic (1PL, 2PL, and 3PL) models were applied to sets of items defined by test section and discipline. The resulting item parameter and examinee score estimates were analyzed in order to address the following questions:

1. What are the relative fits of the 1PL, 2PL, and 3PL models?
2. What is the relative stability across administrations of the parameter estimates based on the 1PL, 2PL, and 3PL models as applied to these data?
3. For the science test sections, how are the item parameter estimates influenced by the possible multidimensionality represented by the disciplines?
4. How are the relative orderings of the examinees influenced by the different test section-level scoring options, including raw scores, scale scores, and scores based on the IRT models?
5. For the science test sections, how are the relative orderings of the examinees influenced by the potential multidimensionality represented by the disciplines?
6. How are the correlations among the test section scores influenced by the choice of test section-level scoring options, including scale scores and the IRT models?
7. For the science test sections, how are the correlations between discipline scores influenced by the choice of IRT models?

ITEM RESPONSE THEORY MODELS

IRT models usually describe the relations of item responses to a hypothetical underlying trait. The investigator may draw conclusions about the underlying trait from an examination of the items and the item parameters generated in the modeling, and from knowledge of the usual use of the test and its value in predicting certain outcomes. For example, the MCAT is intended to predict a student's performance in medical school by measuring his or her Verbal Reasoning skills, as well as her mastery of Biological Sciences and Physical Sciences concepts. To this end, each form of the MCAT contains test sections corresponding to these three areas. The test sections consist of multiple choice items, each of which requires the candidate to apply a scientific concept or demonstrate understanding of a reading passage. The candidate must select the appropriate response for each item from among four alternatives. The investigator applying traditional IRT methods to these test sections would calibrate the items within each test section separately, then examine the item parameters expressing the relations of the items to the underlying trait to verify that those relations are what one would expect if the underlying trait were mastery of the appropriate concepts.

The simplest model used in this study, the one-parameter logistic model, can be written as:

$$P_j(\theta) = 1 / \{1 + \exp [-Da (\theta - b_j)]\}.$$

Here, P is the probability of answering the item correctly at a particular level of θ, the underlying trait or ability. The parameter b_j is related to the difficulty of the item j. The a parameter represents the item's discrimination, but in the 1PL model it is constrained to be the same across all items. D is a scaling constant. If D is 1.7, then a and b have the same values in the normal ogive and logistic versions of the model. The 1PL and 2PL models in MULTILOG set D at 1.0.

The two-parameter logistic model is more flexible, because it allows the slopes of the lines to be different for each item. This model is:

$$P_j(\theta) = 1 / \{1 + \exp [-Da_j (\theta - b_j)]\}.$$

The difference between this model and the previous model is that in this model a has a subscript, j, so that each item can have different discrimination estimates.

A model that allows for a non-zero probability of answering an item correctly even at very low ability levels (which may be appropriate if, for example, examinees are able to guess the correct answer to an item without knowing the answer) is the three-parameter logistic model:

$$P_j(\theta) = c_j + (1 - c_j) / \{1 + \exp [-Da_j (\theta - b_j)]\}.$$

In this equation, c is the probability of getting the question right just by guessing. The c parameter is sometimes referred to as the "guessing parameter," though it can also simply be called the asymptote. The 3PL model in MULTILOG sets D at 1.7.

CALIBRATION AND SCORING PROCEDURES

Estimation of the item parameters for these models involves finding the values for a, b, and, in the 3PL model, c, that yield the highest probability for the observed examinee response patterns. For the analyses reported here, the MULTILOG computer program (Thissen, 1991) was used to compute Marginal Maximum Likelihood (MML) item parameter estimates for the multiple-choice items on three combinations of forms and administrations.

MULTILOG was also used to compute examinee score estimates, based on these item parameter estimates. MULTILOG provides two ways to compute score estimates: (1) Maximum a Priori (MAP) scoring and (2) Maximum Likelihood Estimation (MLE) scoring. MAP scoring is Bayesian: A $N(0,1)$ population prior is incorporated into the estimates. MLE scoring is non-Bayesian: Scores are based on examinees' responses, but may be undefined for examinees with perfectly correct or incorrect response vectors. In this study, MAP scoring was used. For the purposes of these analyses, IRT examinee score estimates were left in the default IRT metric, $N(0,1)$.

In all of the IRT calibration and scoring analyses reported in this chapter, missing data (omitted or not reached items) were counted as "wrong."

ANALYSIS RESULTS

The remainder of this chapter describes the results of the series of analyses addressing the seven questions listed at the beginning of this chapter.

MODEL FIT

The first question, "What are the relative fits of the 1PL, 2PL, and 3PL models?", was addressed by the analyses reported in Table 7. It summarizes the results of the item calibrations for all the items within each test section. The only fit index provided by MULTILOG is negative twice the log likelihood, which has an approximately χ^2 distribution. This index and the model comparisons are included in the table below,. These statistics suggest that the 3PL model fits the data (i.e., is able to reproduce the original response matrices) better than the 2PL model, and that the 2PL model fits the data better than the 1PL model. However, the large sample sizes again make the interpretation of these statistics difficult. Further analyses, reported later in this chapter, provide additional information on which to base comparisons of the models.

TABLE 7. MODEL COMPARISONS FOR MULTILOG ITEM PARAMETER ESTIMATION ASSUMING EACH TEST SECTION TO BE UNIDIMENSIONAL

Form	Test Section	Administration	*n*	Items in Test Section	1PL χ^2	2PL χ^2	3PL χ^2	2PL vs. 1PL $\chi^2(df)$	2PL vs. 1PL *p*	3PL vs. 2PL $\chi^2(df)$	3PL vs. 2PL *p*
X	Verbal Reasoning	1994	16,520	55	656,588	649,953	649,228	6,635(54)	<.001	725(55)	<.001
		1996	3,638	55	150,492	149,256	149,072	1,236(54)	<.001	184(55)	<.001
	Physical Sciences	1994	16,520	63	844,234	836,711	834,196	7,523(62)	<.001	2,515(63)	<.001
		1996	3,638	63	194,273	192,496	191,866	1,777(62)	<.001	630(63)	<.001
	Biological Sciences	1994	16,520	63	845,210	837,358	835,923	7,852(62)	<.001	1,435(63)	<.001
		1996	3,638	63	190,299	188,642	188,303	1,657(62)	<.001	339(63)	<.001
Y	Verbal Reasoning	1996	12,625	55	506,706	499,853	499,277	6,853(54)	<.001	576(55)	<.001
	Physical Sciences	1996	12,625	63	675,501	666,204	664,736	9,297(62)	<.001	1,468(63)	<.001
	Biological Sciences	1996	12,625	63	664,341	659,445	657,776	4,896(62)	<.001	1,669(63)	<.001

Note: 1PL, 2PL, and 3PL = one-, two-, and three-parameter logistic models; the degrees of freedom for the χ^2 statistics for the 1PL, 2PL, and 3PL models are the number of possible response patterns minus the number of parameters estimated minus 1 – i.e., $df = 2^{nitems} - nparameters - 1$; for the model comparisons, the degrees of freedom is simply the difference in the number of parameters estimated; *p* = significance of difference between fits of models.

Examination of the test information curves for Form X in 1994 shown in Figures 1a, 1b, and 1c (the test information curves for Form X in 1996 and Form Y in 1996 are similar) suggests that the 3PL model consistently provides more information in the top half of the score distribution than do the 1PL or 2PL models, while the 1PL model provides more information at the lowest score levels. It may be that the relative importance for medical school admissions decisions of score precision at different points on the score continuum should be a consideration when selecting a model for future calibrations.

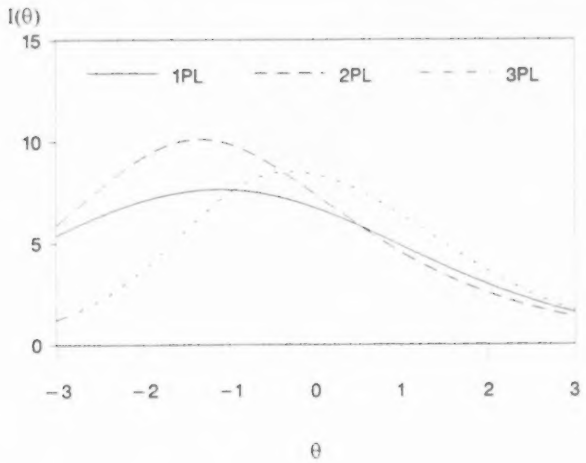

FIGURE 1A. TEST INFORMATION CURVES: FORM X IN 1994, VERBAL REASONING TEST SECTION

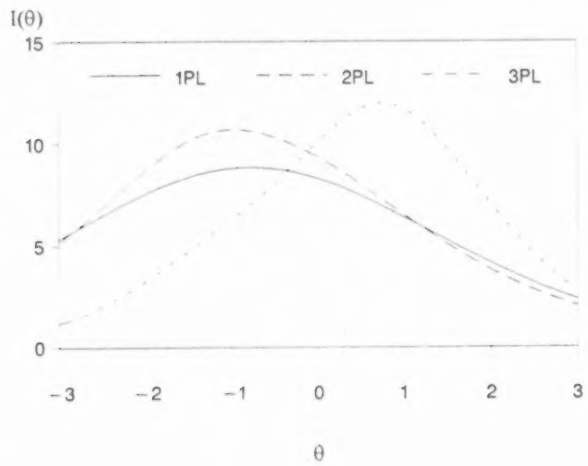

FIGURE 1B. TEST INFORMATION CURVES: FORM X IN 1994, PHYSICAL SCIENCES TEST SECTION

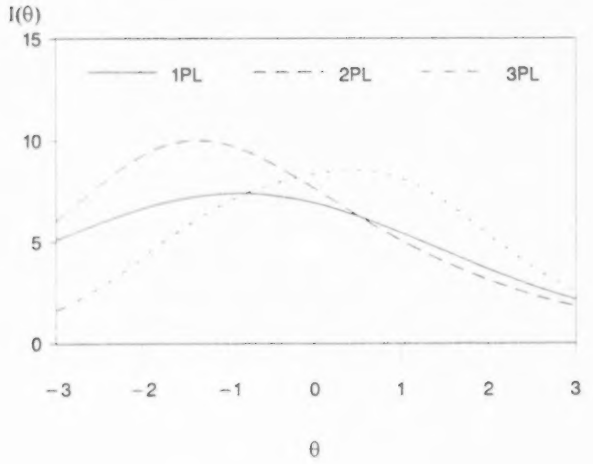

FIGURE 1C. TEST INFORMATION CURVES: FORM X IN 1994, BIOLOGICAL SCIENCES TEST SECTION

ITEM PARAMETER ESTIMATE STABILITY ACROSS ADMINISTRATIONS

The second question posed at the beginning of this chapter was " What is the relative stability across administrations of the parameter estimates based on the 1PL, 2PL, and 3PL models as applied to these data?" To answer this question, the item parameter estimates for the same Form X items administered in 1994 and again in 1996 were compared. Table 8 shows the correlations among the various estimates across administrations. In general, these correlations indicate that the difficulty (*b*) parameters are very stable for all three models, while the slope (*a*) parameters are slightly less so, especially for the 3PL model, and the asymptote (*c*) parameters are the least stable. The greater stability of the difficulty parameters is as would be expected given the model definitions.

TABLE 8. CORRELATIONS AMONG ITEM PARAMETER ESTIMATES: FORM X IN 1994 AND FORM X IN 1996

Model	Parameter	Verbal Reasoning	Physical Sciences	Biological Sciences
1PL	*b*	0.988	0.994	0.995
2PL	*a*	0.949	0.959	0.943
	b	0.977	0.992	0.976
3PL	*a*	0.899	0.913	0.917
	b	0.949	0.963	0.981
	c	0.639	0.806	0.866

Influence of Discipline-Based Multidimensionality on Item Parameter Estimates for Science Test Sections

The third question, "For the science test sections, how are the item parameter estimates influenced by the possible multidimensionality represented by the disciplines?", is addressed by the analyses reported in Table 9.

In addition to the calibrations by test section, the Physical Sciences and Biological Sciences items were also calibrated by discipline within test section. For example, the General Chemistry items and the Physics items on each form of the Physical Sciences test section were separated into two sets and each set was calibrated as though it were a separate test. The item parameter estimates obtained by calibrating the items in this way (i.e., within discipline) were correlated with the item parameter estimates obtained by calibrating all the items on the test section together (i.e., across discipline). Table 9 reports these correlations for each discipline within each science test section.

As one would expect, based on the attributes of the IRT models, the correlations reported in Table 9 are greatest for the difficulty (b) parameters. Also, not surprisingly, the magnitudes of the correlations are related to the proportion of the items within a test section that belong to the discipline. Almost two-thirds of the Biological Sciences items have a discipline designation of Biology, while only one-third are Organic Chemistry items. It makes sense that the Biology discipline items would have a greater influence on defining the underlying factor in the Biological Sciences test section calibration than would the Organic Chemistry items. Consequently, the item parameter estimates for the Biology items calibrated alone are slightly more similar to the estimates for those items when calibrated with all of the Biological Sciences items than are the item parameter estimates for the Organic Chemistry items when the Organic Chemistry items are calibrated alone. The Physical Sciences test section, in contrast, is split almost evenly between the General Chemistry and Physics disciplines; this even split is reflected in the more comparable correlations for the two disciplines.

Generally speaking, however, the results show that the estimates of the IRT parameters for both test sections are relatively unaffected by whether they are estimated within or across disciplines, suggesting that the multidimensionality of the science test sections may be somewhat immaterial in the calibration of the MCAT item bank. If the test sections were strongly multidimensional, one would expect the θ scale for the joint calibration to be defined by only one of the dimensions, so that the items on the other dimension would have strikingly different item parameter estimates on the joint calibration than they had when calibrated alone. The finding that the item parameter estimates for the items of both disciplines within each test section are in fact quite similar, whether calibrated together or separately, suggests that the degree of multidimensionality is small.

TABLE 9. CORRELATIONS AMONG ITEM PARAMETER ESTIMATES CALIBRATED BY TEST SECTION ACROSS DISCIPLINE AND WITHIN DISCIPLINE

Form	Administration	Model	Parameter	Physical Sciences		Biological Sciences	
				General Chemistry	Physics	Organic Chemistry	Biology
X	1994	1PL	*b*	1.000	1.000	1.000	1.000
		2PL	*a*	0.984	0.996	0.915	0.996
			b	0.998	1.000	0.977	1.000
		3PL	*a*	0.976	0.976	0.902	0.990
			b	0.984	0.996	0.980	0.999
			c	0.834	0.963	0.852	0.990
X	1996	1PL	*b*	1.000	1.000	1.000	1.000
		2PL	*a*	0.984	0.988	0.906	0.994
			b	0.999	0.999	0.968	1.000
		3PL	*a*	0.980	0.971	0.878	0.957
			b	0.993	0.996	0.972	0.998
			c	0.944	0.964	0.815	0.965
Y	1996	1PL	*b*	1.000	1.000	1.000	1.000
		2PL	*a*	0.994	0.991	0.922	0.993
			b	0.999	0.998	0.995	0.998
		3PL	*a*	0.961	0.938	0.927	0.964
			b	0.995	0.994	0.984	0.994
			c	0.979	0.966	0.778	0.974

Note: Correlations presented are for the items in a discipline (General Chemistry, Physics, Organic Chemistry, or Biology), between the item parameter estimates from calibration of only the items in that discipline and those yielded from calibration of all the items in a test section (Physical Sciences or Biological Sciences) together.

SCORE COMPARABILITY WITHIN TEST SECTION

Based on the various item calibrations, score estimates for the examinees in the example data sets were computed. Using these score estimates, a series of analyses were performed to address the questions about score comparability (i.e., questions 4 through 7).

The fourth question is "How are the relative orderings of the examinees influenced by the different test section-level scoring options, including raw scores, scale scores, and scores based on the IRT models?" Correlations among the various scores -- raw scores, scale scores (which are based on the operational raw-to-scale score conversion tables and which are the scores reported to examinees), and the θ scores based on 1PL, 2PL, and 3PL IRT calibrations -- are presented in Tables 10a through 10c for the Form X 1994 and 1996 data, and Tables 11a through 11c for the Form Y 1996 data.

As one might expect, the correlations among the scores presented in these tables are all very high. Of the IRT models, the 1PL model consistently yields scores with the greatest correlations with the original raw and scale scores. This is as expected because the 1PL, like the raw score on which the scale score is based, but unlike the 2PL and 3PL models, does not differentiate among items in terms of their discrimination. Thus, all items contribute equally to the score estimate.

TABLE 10A. CORRELATIONS AMONG RAW, SCALE, AND IRT-BASED SCORES: FORM X IN 1994 AND 1996, VERBAL REASONING TEST SECTION

	Raw		Scale		IRT: 1PL		IRT: 2PL		IRT: 3PL	
	1994	1996	1994	1996	1994	1996	1994	1996	1994	1996
Raw	1.000	1.000								
Scale	0.992	0.992	1.000	1.000						
IRT: 1PL	0.995	0.994	0.989	0.989	1.000	1.000				
IRT: 2PL	0.985	0.985	0.981	0.980	0.993	0.994	1.000	1.000		
IRT: 3PL	0.988	0.988	0.982	0.982	0.992	0.992	0.998	0.997	1.000	1.000

TABLE 10B. CORRELATIONS AMONG RAW, SCALE, AND IRT-BASED SCORES: FORM X IN 1994 AND 1996, PHYSICAL SCIENCES TEST SECTION

	Raw		Scale		IRT: 1PL		IRT: 2PL		IRT: 3PL	
	1994	1996	1994	1996	1994	1996	1994	1996	1994	1996
Raw	1.000	1.000								
Scale	0.988	0.988	1.000	1.000						
IRT: 1PL	0.995	0.994	0.991	0.991	1.000	1.000				
IRT: 2PL	0.989	0.988	0.985	0.985	0.995	0.995	1.000	1.000		
IRT: 3PL	0.989	0.987	0.980	0.979	0.987	0.985	0.992	0.991	1.000	1.000

TABLE 10C. CORRELATIONS AMONG RAW, SCALE, AND IRT-BASED SCORES: FORM X IN 1994 AND 1996, BIOLOGICAL SCIENCES TEST SECTION

	Raw		Scale		IRT: 1PL		IRT: 2PL		IRT: 3PL	
	1994	1996	1994	1996	1994	1996	1994	1996	1994	1996
Raw	1.000	1.000								
Scale	0.991	0.991	1.000	1.000						
IRT: 1PL	0.997	0.996	0.989	0.989	1.000	1.000				
IRT: 2PL	0.988	0.987	0.979	0.981	0.993	0.994	1.000	1.000		
IRT: 3PL	0.990	0.990	0.982	0.983	0.990	0.990	0.996	0.995	1.000	1.000

TABLE 11A. CORRELATIONS AMONG RAW, SCALE, AND IRT-BASED SCORES: FORM Y IN 1996, VERBAL REASONING TEST SECTION

	Raw	Scale	IRT: 1PL	IRT: 2PL	IRT: 3PL
Raw	1.000				
Scale	0.990	1.000			
IRT: 1PL	0.998	0.990	1.000		
IRT: 2PL	0.983	0.979	0.988	1.000	
IRT: 3PL	0.984	0.977	0.986	0.997	1.000

TABLE 11B. CORRELATIONS AMONG RAW, SCALE, AND IRT-BASED SCORES: FORM Y IN 1996, PHYSICAL SCIENCES TEST SECTION

	Raw	Scale	IRT: 1PL	IRT: 2PL	IRT: 3PL
Raw	1.000				
Scale	0.985	1.000			
IRT: 1PL	0.996	0.992	1.000		
IRT: 2PL	0.982	0.980	0.987	1.000	
IRT: 3PL	0.983	0.975	0.984	0.995	1.000

TABLE 11C. CORRELATIONS AMONG RAW, SCALE, AND IRT-BASED SCORES: FORM Y IN 1996, BIOLOGICAL SCIENCES TEST SECTION

	Raw	Scale	IRT: 1PL	IRT: 2PL	IRT: 3PL
Raw	1.000				
Scale	0.992	1.000			
IRT: 1PL	0.996	0.986	1.000		
IRT: 2PL	0.989	0.979	0.994	1.000	
IRT: 3PL	0.989	0.982	0.989	0.993	1.000

Figures 2a, 2b, and 2c illustrate the relation between the raw scores and IRT scores for the Biological Sciences Test Section of Form X in 1994. The plots for the other test sections and other forms and administrations are similar.

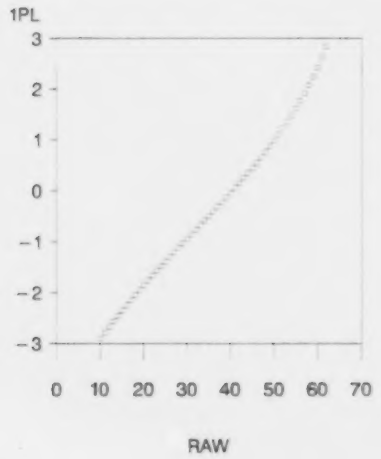

FIGURE 2A. EXAMINEE SCORE ESTIMATE SCATTER PLOTS: FORM X, 1994, BIOLOGICAL SCIENCES TEST SECTION, 1-PARAMETER LOGISTIC MODEL AND RAW SCORES

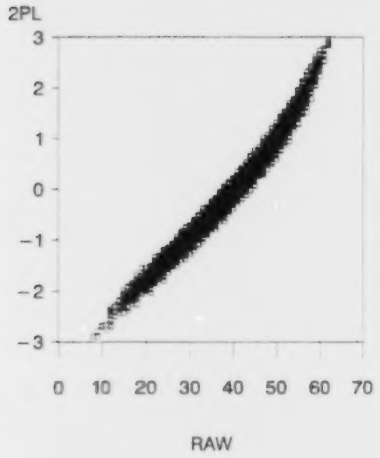

FIGURE 2B. EXAMINEE SCORE ESTIMATE SCATTER PLOTS: FORM X, 1994, BIOLOGICAL SCIENCES TEST SECTION, 2-PARAMETER LOGISTIC MODEL AND RAW SCORES

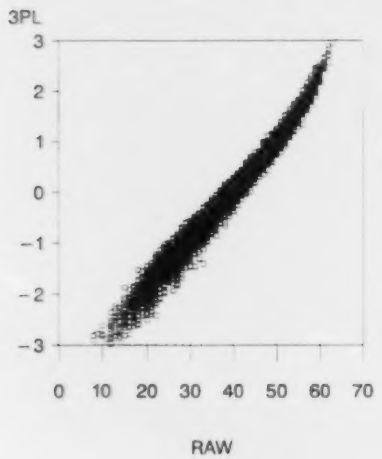

FIGURE 2C. EXAMINEE SCORE ESTIMATE SCATTER PLOTS: FORM X, 1994, BIOLOGICAL SCIENCES TEST SECTION, 3-PARAMETER LOGISTIC MODEL AND RAW SCORES

Influence of Discipline-Based Multidimensionality on Score Estimates for Science Test Sections

The fifth question asks "For the science test sections, how are the relative orderings of the examinees influenced by the potential multidimensionality represented by the disciplines?" To address this question, Physical Sciences and Biological Sciences test section scores were computed for each examinee based on both the within- and across-discipline item parameters described previously. In the former case, scores were computed by discipline, and then a weighted average was computed, with weights based on the target ratio of items in each discipline. Correlations between the two sets of scores are reported in Table 12. The results for both the Physical Sciences and Biological Sciences test sections indicate that the relative ordering of examinees produced by combining the separate scores based on the calibration of the items within discipline is almost identical to the ordering based on a single calibration by test section across discipline. For the data sets examined here, each of the correlations are greater than .99.

Table 12. Correlations Among IRT-Based Score Estimates for Science Test Section Scores Based on Within- and Across-Discipline Item Calibrations

Form	Administration	Model	Physical Sciences	Biological Sciences
X	1994	1PL	0.999	0.998
		2PL	0.999	0.998
		3PL	0.998	0.999
X	1996	1PL	0.999	0.995
		2PL	0.999	0.998
		3PL	0.998	0.998
Y	1996	1PL	0.999	0.996
		2PL	0.998	0.997
		3PL	0.998	0.992

Note: Within-discipline test section scores are weighted averages of scores computed for each discipline based on the separate calibration of the items in each discipline. The weights are based on the target numbers of items of each discipline type.

Score Comparability Across Test Sections

The sixth question addressed in these analyses is "How are the correlations among the test section scores influenced by the choice of test section-level scoring options, including scale scores and the IRT models?" Table 13 reports these correlations for each of the three data sets. These results show that the pattern and level of correlations among the three test sections are very similar between the scale scores and the three IRT-based score estimates. This suggests that the factor structure of the MCAT is not significantly affected by the choice of scoring models.

TABLE 13. CORRELATIONS AMONG TEST SECTIONS FOR SCALE AND IRT-BASED SCORE ESTIMATES

Form	Administration	Model/Score	Verbal Reasoning with Physical Sciences	Verbal Reasoning with Biological Sciences	Physical Sciences with Biological Sciences
X	1994	Scale Score	0.568	0.609	0.783
		1PL Model	0.574	0.609	0.800
		2PL Model	0.570	0.608	0.804
		3PL Model	0.586	0.618	0.811
X	1996	Scale Score	0.559	0.598	0.786
		1PL Model	0.566	0.601	0.804
		2PL Model	0.561	0.599	0.807
		3PL Model	0.574	0.607	0.814
Y	1996	Scale Score	0.547	0.625	0.783
		1PL Model	0.564	0.629	0.803
		2PL Model	0.542	0.633	0.804
		3PL Model	0.552	0.640	0.807

SCORE COMPARABILITY ACROSS DISCIPLINES

Finally, Table 14 contains information relevant to the last question, "For the science test sections, how are the correlations between discipline scores influenced by the choice of IRT models?" The correlations presented in Table 14 are between the discipline-level scores for each test section based on the calibrations of the items within each discipline. These results indicate that the pattern and level of correlations are very similar across the three models, suggesting that the internal structure of the two science test sections are not significantly influenced by the choice of IRT models used to score examinees.

TABLE 14. CORRELATIONS AMONG DISCIPLINES FOR IRT-BASED SCORE ESTIMATES

Form	Administration	Model	General Chemistry with Physics	Organic Chemistry with Biology
X	1994	1PL	0.737	0.604
		2PL	0.750	0.598
		3PL	0.747	0.604
X	1996	1PL	0.744	0.615
		2PL	0.761	0.612
		3PL	0.756	0.617
Y	1996	1PL	0.703	0.580
		2PL	0.726	0.593
		3PL	0.730	0.597

SUMMARY

The results of the IRT analyses described in this chapter suggest the following answers to the questions posed at the beginning of the chapter:

1. What are the relative fits of the 1PL, 2PL, and 3PL models?

   The 3PL model fits the data slightly better than the 1PL or 2PL models and provides greater score precision at the higher scores. In contrast, the 1PL model provides greater precision at the lower scores.

2. What is the relative stability across administrations of the parameter estimates based on the 1PL, 2PL, and 3PL models as applied to these data?

   The difficulty (b) parameters are very stable, while the slope (a) parameters are slightly less so, and the asymptote (c) parameters are the least stable. The greater stability of the difficulty parameters is as would be expected given the model definitions. The correlations among the difficulty parameters across administrations range from .94 to .99; among the slope parameters from .90 to .96.

3. For the science test sections, how are the item parameter estimates influenced by the possible multidimensionality represented by the disciplines?

   The estimates of the IRT parameters are relatively unaffected by whether they are estimated within or across disciplines, suggesting that the discipline-based multidimensionality of the science test sections may be somewhat immaterial in the calibration of the MCAT item bank. The correlations among the difficulty parameters are all above .97, and among the slope parameters above .88.

4. How are the relative orderings of the examinees influenced by the different test section-level scoring options, including raw scores, scale scores, and scores based on the IRT models?

   The correlations among the various types of test section scores are all very high (above .97). Of the IRT models, the 1PL model consistently yields scores with the highest correlations with the original raw and scale scores. This is as expected, because the 1PL, like the raw score on which the scale score is based (but unlike the 2PL and 3PL models), does not differentiate among items in terms of their discrimination. Thus, all items contribute equally to the score estimate.

5. For the science test sections, how are the relative orderings of the examinees influenced by the potential multidimensionality represented by the disciplines?

   For both the Physical Sciences and Biological Sciences test sections, the relative ordering of examinees produced by combining the separate scores based on the calibration of the items by discipline is almost identical to the ordering based on a single calibration by test section. For all of the data sets, these correlations are greater than .99.

6. How are the correlations among the test section scores influenced by the choice of test section-level scoring options, including scale scores and the IRT models?

   The pattern and level of correlations are very similar across IRT models and between the IRT model scores and the scale scores (about .55 for Verbal Reasoning with Physical Sciences, .60 for Verbal Reasoning with Biological Sciences, and .80 for Physical Sciences with Biological Sciences). These correlations are consistent with those found for the typical MCAT administration.

7. For the science test sections, how are the correlations between discipline scores influenced by the choice of IRT models?

   The pattern and levels of correlations are very similar across models (.70 to .76 for General Chemistry with Physics and .58 to .62 for Organic Chemistry with Biology).

The implications of these findings for future IRT analyses of MCAT data are discussed in the last chapter.

4. TREATMENT OF MISSING DATA

In the course of conducting this study, two sets of analyses were performed. The analyses were parallel, differing only in the treatment of missing data. In the IRT calibration and scoring analyses reported in the previous chapter, missing data (omitted or not reached items) were scored as "wrong." In the other set of analyses, the results of which are not presented in detail in this report, omitted and not reached items were both ignored -- that is, the calibration and scoring computations were based only on examinees' responses.

A comparison of these two sets of results is informative about the impact of the treatment of missing data on item parameter estimates and examinee score estimates. Of course, it is important to note that these treatments of missing data are only two of many possibilities. Other variations might include distinguishing between omitted and not reached items, counting missing items as fractionally correct, and treating missing items differently in the calibration and in the scoring analyses.

This chapter provides three tables comparing the results of the two sets of analyses. The first of these, Table 15, presents the numbers of items omitted or not reached by examinees. As this table shows, the percentage of examinees answering all the items in a test section ranged from 85 percent for the Physical Sciences test section on Form X administered in 1994 to almost 96 percent for the Biological Sciences test section on Form X in 1996. For each form and administration, the percentage of examinees answering all the items was consistently highest for the Biological Sciences test section, which is administered last. As Table 15 shows, a few examinees did omit large numbers of items. However, these examinees represent a small percentage of the examinee population.

TABLE 15. NUMBERS OF ITEMS OMITTED OR NOT REACHED

Form	Administration	Number of Omitted or Not Reached Items	Number (Percent) of Examinees		
			Verbal Reasoning	Physical Sciences	Biological Sciences
X	1994	0	14,138 (85.6)	14,035 (85.0)	15,694 (95.0)
		1	1,132 (6.9)	1,576 (9.5)	655 (4.0)
		2-5	755 (4.6)	642 (3.9)	136 (0.8)
		6-10	281 (1.7)	155 (0.9)	22 (0.1)
		11-15	133 (0.8)	60 (0.4)	3 (0.0)
		16-20	62 (0.4)	29 (0.2)	6 (0.0)
		21-25	11 (0.1)	16 (0.1)	3 (0.0)
		26-30	5 (0.0)	3 (0.0)	0 (0.0)
		>30	3 (0.0)	4 (0.0)	1 (0.0)
X	1996	0	3,184 (87.5)	3,124 (85.9)	3,487 (95.8)
		1	204 (5.6)	314 (8.6)	121 (3.3)
		2-5	155 (4.3)	124 (3.4)	24 (0.7)
		6-10	66 (1.8)	45 (1.2)	4 (0.1)
		11-15	18 (0.5)	15 (0.4)	2 (0.1)
		16-20	9 (0.2)	8 (0.2)	0 (0.0)
		21-25	2 (0.1)	5 (0.1)	0 (0.0)
		26-30	0 (0.0)	2 (0.1)	0 (0.0)
		>30	0 (0.0)	1 (0.0)	0 (0.0)
Y	1996	0	11,171 (88.5)	11,378 (90.1)	11,663 (92.4)
		1	654 (5.2)	831 (6.6)	712 (5.6)
		2-5	480 (3.8)	296 (2.3)	203 (1.6)
		6-10	200 (1.6)	66 (0.5)	30 (0.2)
		11-15	62 (0.5)	28 (0.2)	10 (0.1)
		16-20	33 (0.3)	14 (0.1)	6 (0.0)
		21-25	7 (0.1)	8 (0.1)	1 (0.0)
		26-30	18 (0.1)	3 (0.0)	0 (0.0)
		>30	0 (0.0)	1 (0.0)	0 (0.0)

Table 16 presents the correlations for the item parameter estimates computed with missing data ignored and with missing data scored as wrong. All of the correlations exceed .98, indicating that ignoring missing items rather than scoring them as wrong had very little impact on the calibration results.

TABLE 16. CORRELATIONS AMONG ITEM PARAMETER ESTIMATES ACROSS METHOD OF SCORING MISSING DATA: MISSING ITEMS IGNORED AND MISSING ITEMS SCORED AS WRONG

Form	Administration	Model	Parameter	Verbal Reasoning Test Section	Physical Sciences Test Section	Biological Sciences Test Section
X	1994	1PL	*b*	0.9995	0.9999	1.0000
		2PL	*a*	0.9963	0.9996	1.0000
			b	0.9989	0.9998	1.0000
		3PL	*a*	0.9943	0.9994	0.9999
			b	0.9988	0.9998	1.0000
			c	0.9940	0.9991	0.9998
X	1996	1PL	*b*	0.9995	0.9998	1.0000
		2PL	*a*	0.9977	0.9988	1.0000
			b	0.9988	0.9996	1.0000
		3PL	*a*	0.9806	0.9975	0.9998
			b	0.9972	0.9993	1.0000
			c	0.9735	0.9957	0.9998
Y	1996	1PL	*b*	0.9998	0.9998	1.0000
		2PL	*a*	0.9980	0.9998	0.9999
			b	0.9992	0.9996	0.9999
		3PL	*a*	0.9963	0.9991	0.9999
			b	0.9996	0.9989	0.9998
			c	0.9882	0.9982	0.9993

Note: Correlations presented are for the items in a discipline (General Chemistry, Physics, Organic Chemistry, or Biology), between the item parameter estimates from calibration of only the items in that discipline and those yielded from calibration of all the items in a test section (Physical Sciences or Biological Sciences) together.

Table 17 presents correlations for the examinee score estimates and shows that the examinee score estimates computed under the two missing data treatment conditions are very similar, with all correlations exceeding .99. These high correlations are not surprising, given the small percent of examinees that omitted items.

TABLE 17. CORRELATIONS AMONG IRT-BASED SCORE ESTIMATES ACROSS METHOD OF SCORING MISSING DATA: MISSING ITEMS IGNORED AND MISSING ITEMS SCORED AS WRONG

Form	Administration	Model	Verbal Reasoning	Physical Sciences	Biological Sciences
X	1994	1PL	0.9912	0.9962	0.9995
		2PL	0.9914	0.9968	0.9996
		3PL	0.9916	0.9969	0.9996
X	1996	1PL	0.9912	0.9958	0.9997
		2PL	0.9911	0.9966	0.9997
		3PL	0.9908	0.9965	0.9998
Y	1996	1PL	0.9912	0.9975	0.9992
		2PL	0.9922	0.9981	0.9992
		3PL	0.9924	0.9981	0.9992

Figures 3a, 3b, and 3c provide scatter plots for three examples of the correlations in Table 17 (the plots for the other test sections, forms, and administrations are similar). These plots show that the examinees who did not answer every item received higher scores when these missing items were ignored than when they were scored as wrong. However, as the high correlations in Table 17 suggest, the proportion of examinees with large score differences is small.

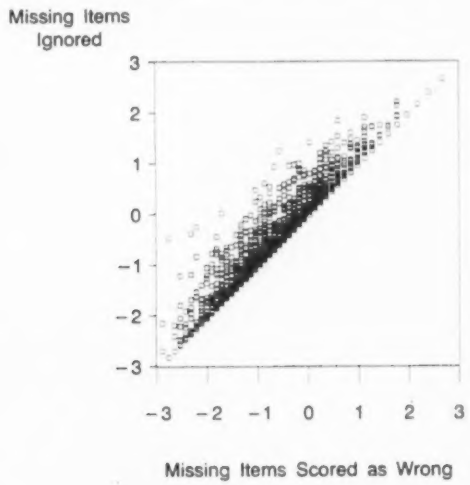

FIGURE 3A. EXAMINEE SCORE ESTIMATE SCATTER PLOTS: FORM X, 1994, VERBAL REASONING TEST SECTION, 1-PARAMETER LOGISTIC MODEL, MISSING ITEMS IGNORED AND MISSING ITEMS SCORED AS WRONG

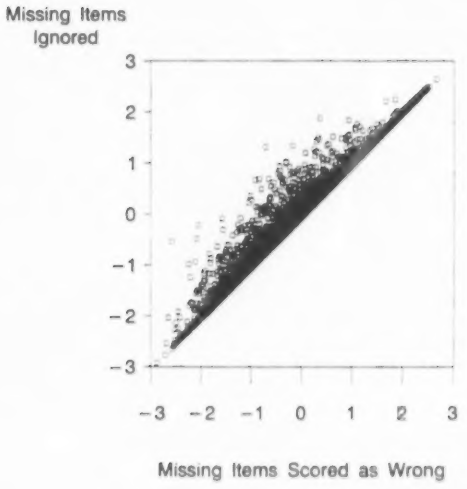

FIGURE 3B. EXAMINEE SCORE ESTIMATE SCATTER PLOTS: FORM X, 1994, VERBAL REASONING TEST SECTION, 2-PARAMETER LOGISTIC MODEL, MISSING ITEMS IGNORED AND MISSING ITEMS SCORED AS WRONG

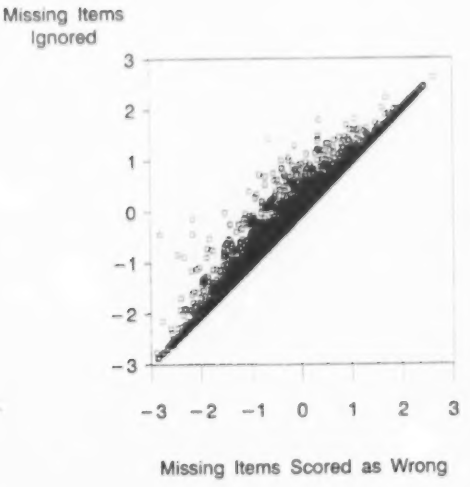

FIGURE 3C. EXAMINEE SCORE ESTIMATE SCATTER PLOTS: FORM X, 1994, VERBAL REASONING TEST SECTION, 3-PARAMETER LOGISTIC MODEL, MISSING ITEMS IGNORED AND MISSING ITEMS SCORED AS WRONG

As the results presented in this chapter show, treating missing items as wrong or ignoring them does result in differences in calibration and scoring results. These differences are small, most likely because most of the examinees taking the MCAT answer every item. However, for the small number of examinees who do omit large numbers of items, their score estimates are likely to be greatly impacted by the decision of how to treat missing data -- whether using one of these approaches or some other approach. The decision of how to treat missing data in possible future IRT analyses of the MCAT may also be based on comparability with current practice, among other considerations.

5. CONCLUSIONS

The dimensionality analyses described in this report suggest that, while the items within each of the science test sections are not completely homogeneous, neither are they clearly measuring distinct constructs corresponding to the disciplines. While the Biological Sciences test section shows some evidence for two factors corresponding to the two discipline types, the Physical Sciences test section does not appear to support two distinguishable factors. When confirmatory factor analyses are performed on the test sections, the disciplines are correlated about .85 for the Biological Sciences test section and about .95 for the Physical Sciences test section.

This is not to say that other dimensions (possibly related to content areas, for example,) do not exist in the data. These analyses do not show the data to be unidimensional -- although the large sample sizes may have caused some of these indices to be overly sensitive to small deviations from unidimensionality. However, neither do the analyses provide strong support for the factors hypothesized.

The motivation for performing these analyses was to determine whether the test sections deviated sufficiently from unidimensionality to require that IRT item calibrations be performed separately by discipline within test section. These results suggest that that they do not. Indeed, comparisons of the results of IRT calibration and scoring of the data by test section and by discipline within test section provide additional support for this conclusion. Specifically, the correlation between IRT scores based on item parameters that were estimated separately within discipline and then formed into weighted composites and scores based on item parameters that were estimated across discipline (within test section) exceed .99.

An additional issue addressed by these analyses concerns the question of which IRT model is most appropriate for calibrating the items in the MCAT item bank. Once again, the relatively high correlations among the scores based on the various IRT models and between those scores and the non-IRT-based raw and scale scores indicate that the 1PL, 2PL, and 3PL models would each yield extremely similar results with regard to the relative ordering of MCAT examinees on the three multiple-choice test sections. This suggests that the ultimate decision regarding the choice of IRT models to implement may be based on other considerations, such as the likely defensibility of pattern scoring or the complexity of the calibration analyses. It should be noted, however, that one difference among the models that may merit some attention concerns the test information curves associated with the different models. Specifically, the results of these analyses suggest that the 3PL models tend to yield more information for examinees with score estimates above the midpoint of the score continuum, while the 1PL models tend to provide more information for examinees with score estimates in the bottom half of the continuum.

The use of these results to inform possible future IRT analyses of the MCAT depends in large part on policy decisions about future directions of that program. However, these results may also have implications for other testing programs as those programs seek to determine the practical implications of dimensionality for the IRT calibration of their tests.

REFERENCES

Bejar, I. I. (1980). "A procedure for investigating the unidimensionality of achievement based on item parameter estimates." *Journal of Educational Measurement, 17,* 282-296.

DeAyala, R. J., & Hertzog, M. A. (1991). "The assessment of unidimensionality for use in item response theory." *Multivariate Behavioral Research, 26,* 765-792.

De Champlain, A. F., & Gessaroli, M. E. (in press). "Assessing the dimensionality of item response matrices with small sample sizes and short test lengths." *Applied Measurement in Education.*

De Champlain, A. F., & Tang, K. L. (1997). "CHIDIM: A FORTRAN program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model." *Educational and Psychological Measurement, 57,* 174-178.

Fraser, C. (1988). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, New South Wales, Australia: Center for Behavioral Studies, University of New England.

Gessaroli, M. E., & De Champlain, A. F. (1996). "Using an approximate chi-square statistic to test the number of dimensions underlying responses to a set of items." *Journal of Educational Measurement, 33,* 157-179.

Hambleton, R. K., & Rovenelli, R. J. (1986). "Assessing the dimensionality of a set of test items." *Applied Psychological Measurement, 10,* 287-302.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). "An assessment of Stout's index of essential unidimensionality." *Applied Psychological Measurement, 20,* 1-14.

Jöreskog, K. G., & Sörbom, D. (1988). *PRELIS - A program for multivariate data screening and data summarization. A preprocessor for LISREL* (2nd ed.). Chicago, IL: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1993a). *LISREL 8: Structural equation modeling with the SIMPLIS command language* . Chicago, IL: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1993b). *New features in PRELIS 2.* Chicago, IL: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1993c). *New features in LISREL 8.* Chicago, IL: Scientific Software.

McDonald, R. P. (1994). "Testing for approximate dimensionality." In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern Theories of Measurement: Problems and issues* (pp. 63-86). Ottawa: Edumetrics Research Group.

Mislevy, R. J. (1986). "Recent developments in the factor analysis of categorical variables." *Journal of Educational Statistics, 11,* 3-31.

Nandakumar, R. (1994). "Assessing dimensionality of a set of item responses—comparison of different approaches." *Journal of Educational Measurement, 31,* 17-35.

Sireci, S. G. (1997). *Dimensionality issues related to the National Assessment of Educational Progress.* Commissioned paper by the National Academy of Sciences/National Research Council's Committee on the Evaluation of National and State Assessments of Educational Progress.

Stout, W. (1987). "A nonparametric approach for assessing latent trait unidimensionality." *Psychometrika,* 52, 589-617.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). *DIMTEST manual.* Champaign, IL: Department of Statistics, University of Illinois at Urbana-Champaign.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using Item Response Theory* (Version 6.0). Chicago: Scientific Software.